# A general format for time information to be the first-class data of general linguistics

**Kazushi Ohya**
**Tsurumi University**

2015-02-27

A shared data format for time information is needed for linguists to use sound data as the first-class data as well as encoded language data. It is better if this format is a plain text in a flat data model and each record can be in super-set order. This is based on a policy that convertibility ensures preservability. We temporarily call this format GIST; a format of general information of sub-time for linguistics. This data format can be used to help to realize multi-layered objects observed in language phenomena.

## 1 Summary

This presentation aims to propose a philosophy and an actual description of data format for time information, to realize a new phenomenon language documentation brings about with computational environments. A data format for time information is needed to make sound data primary data units as well as encoded and annotated language resources. This data format should be simple and cogent because it is used by linguists as a common and fundamental format to record sound data to relate it to encoded language data. To be simple, the data format is based on a flat data model and the actual description is better in a plain text. To be cogent, the data format is based on mathematical foundations. This kind of simple format is compliant with a policy that convertibility ensures preservability. The simplicity will contribute both for making and preserving language data. In this proposal, elements in records line up in superset order, which means that a left-side element is a superset of the right-side element. The elements are an ID or an equivalent of it, such as a file name or a pair of time information with start and end timestamps. An example of the actual description of a record is "original_sound.wav,00:00:13,00:01:03.25,part_of_sound1.wav." The format proposed in this presentation is not the result of a large academic study, but could be a case example for linguists to start considering the need for time information in their language documentation.

## 2 Background

The reasons why this type of data format is needed are as follows.

(1) As shown in [9] a key strategy for sharing language resources is a data conversion service. From our experiments, data formats based on a multi-link-path model proposed by international organizations or research projects such as IOS LAF/GrAF[2, 6] and TEI[1, 4] have drawbacks of data size and data manipulation[7]. Multiple link paths in a data model require us to know a definite query direction of link paths, steps of link traversal, addressing formats, terminal nodes, and others in advance, and make data size bigger than plain marked-up texts. If we use these formats we have to prepare flexible data conversion programs or services [8, 9]. In order to reduce the number of link paths, a part defining data units in a stand-off style can be moved and be an independent data file. The data format proposed here can be used for this kind of data.

(2) There is a one-to-many relationship between actual sound and sound data, and a many-to-many relationship between sound data and encoded language data. Thus, when replaying sound from sound data, which means realizing the sound in reality again, there must be at least time information as a part of information to realize it. If we need to realize it more precisely, there must be information about environmental conditions such as performance, presentation, or staging. To realize sound data as the first-class data in linguistics, linguists have to make information about this relationship, which implies that linguists have to change their position from consumers to providers of sound data psychologically and practically. This data format proposed here can be used to indicate the relationship. This kind of data format can be regarded as being in one layer of language phenomena based on an idea of "multiple articulation" that is an extended idea of "double articulation" in linguistics.

## 3 Definition of GIST format

The format for time information proposed in this presentation is tentatively called GIST; a format of general information of sub-time for linguistics. This format is a set of records consisting of identifiers of time-based objects in super-set order. Each identifier of time-based objects is a super or sub element of the adjacent ele-

This paper is a handout for ICLDC4 in Hawaii, February 27, 2015.

ments, thus this format can be regarded as for indicating sub-time information. The identifiers are an ID or an equivalent of it such as a file name, or a pair of time information with start and end time-stamp.

### 3.1 Syntax

A syntax of GIST is defined as follows.

$$
\begin{aligned}
&GIST := I+;\\
&I := (N|T)+;\\
&N := NAME;\\
&T := (TIME, TIME);\\
&TIME := hh : mm : ss([.,]d+)?;\\
&NAME := \{\text{any strings}\};
\end{aligned}
$$

N and T are identifiers(I) of time-based objects. N is a name and T is a pair of time-stamps represented by [hh]:[mm]:[ss] style that is similar to a part of times in ISO 8601 format[3], and [ss] can be extended with a comma or dot and decimal fractions. A name for N is a string of characters for the time being.

### 3.2 Semantics

Semantics of GIST is simple: a left object is a super object of a right object in a sequence of objects. An object is indicated by an identifier, which can be a name(N) or a pair of time-stamps(T). Given an object is denoted by an identifier(I), semantics of GIST is defined as follows.

$$[\![I_1 I_2]\!] := I_1 \supseteq I_2$$

If I is expanded into N or T, expressions with a name and a pair of time-stamps are as follows.

$$
\begin{aligned}
&[\![N_1 N_2]\!] := N_1 \supseteq N_2\\
&[\![NT]\!] := N \supseteq T\\
&[\![TN]\!] := T \supseteq N\\
&[\![T_1 T_2]\!] := T_1 \supseteq T_2
\end{aligned}
$$

### 3.3 Semantics in implementation

A relation of superset order is interpreted in actual implementation as follows.

$NT$

N is a name of an object, which is a domain of an object indicated by T that is time information. If there is no N, the domain of T is explicitly given in implementation.

$N_1 T N_2$

$N_1$ is a name of an object, which is a domain of an object indicated by T. The object portioned by T is named by $N_2$, which means that $N_2$ is regarded as a name of the same object as that of T. If there is no $N_1$, the domain of T is explicitly given in implementation.

$NT_1T_2$

N is a name of an object, which is a domain of an object indicated by $T_1$, and the object made by $T_1$ is also a domain of an object indicated by $T_2$. If there is no N, the domain of $T_1$ is explicitly indicated in implementation.

$N_1 N_2$

N is a name of an object, which is a super-set of an object named $N_2$. In implementation, the object named $N_2$ can be regarded as the same object named $N_1$ in implementation. However, in a strict sense, the sameness is ensured with existence of another definition of $N_2 N_1$.

## 4 Sample Implementations

As sample applications using this GIST format, we made software Sclip(sound clip) that cuts out a part of the sound according to instructions in a GIST format. Provided the following instructions and a shared sound data file are there,

```
00:00:01.2,00:00:50.3,file1.wav
00:00:01.2,00:00:50.3,00:00:00,00:00:15,file2.wav
00:00:01.2,00:00:50.3
00:00:01.2,00:00:50.3,00:00:00,00:00:15
00:00:02.2,00:00:50.3,file4.wav,00:00:01.2,00:00:10.32
```

the sclip makes a new sound file named file1.wav from the shared sound file according to the first record, and a new sound data of file2.wav from 00:00:01.2 to 00:00:16.2 in the shared sound data according to the second record. If there is no name at the final item in a record, the sclip defines a file name automatically and saves a result of partial sound data. The functions of extracting a part of sound from a sound object and giving a name to a part of sound are iteratively activated while corresponding to rules in GIST.

As another sample application, we made a simple and easy HTML file to show FLEx data with sound data and time information made by ELAN. This HTML file uses an XML data exported from FLEx and an EAF file made by ELAN that imports the XML data as a base text data then adds region information with a sound file. This HTML file provides a function to play a part of the sound corresponding to an annotation with a click of a mouse. As a support application of this HTML file, we made another html file to distill IDs from FLEx data for EAF files.

## 5 Problems of Name Resolution

If we seek, in language documentation, a common way of handling language data in a computational environment from fieldwork through linguistic analysis to publication, instead of using the same software, we need a system for handling name information appearing in any steps of documentation.

For example, an annotated text has a name for its file, and names for their components such as stories, sentences, or morphemes. FLEx can store IDs of data units in multiple layers. Sound data also has a name for its file, and names for a part of it. ELAN can store IDs of sound units in allocated regions. In addition, there may be a name for a data unit to connect the annotated text and the sound data. For example, ELAN connects annotated texts made by FLEx with sound selected in ELAN by adding a suffix with the same ID used in FLEx. But, if we use software other than ELAN, we need a data unit to indicate this relation, which must have its own name. In a GIST format, a part of sound can be anonymous and be given its name. Thus, data in GIST can be regarded as metadata for defining parts of data units and relations of them. It means that we need another metadata(hub-metadata) describing connections of metadata(GIST), annotated text, and sound data.

This chain of names is similar to the successor operation in Peano axioms, or unlimited. To solve this name resolution problem, we have to adopt a way of seeking a trade-off, which is possible after defining the outer rim of data usage. In our projects, we restrict the domain of data usage within from distilling data from field notes to show sound and annotated texts of languages in Siberia.

## 6 Our Projects

We have studied language documentation and sought new systems and tools used for it in the past two projects: LingDy Project supported by Tokyo University of Foreign Studies[10] from 2008 to 2010 and A Study of Digital Archive Environment and Language Documentation for Minority Languages in North-East Eurasia supported by Grants-in-Aid for Scientific Research in Japan from 2011 to 2014. In these past projects, we found out practical and theoretical problems in data formats[9]. On the basis of the results, we started a new three-year project: A Study of Documentation on Theories and Practices on Minority Languages in Siberia supported by Grants-in-Aid for Scientific Research in Japan from 2014. This project is a fundamental study for a data format to connect time information to encoded language resources by using language data in Siberia as test resources. As we have observed, handling sound data requires us to encode environmental information as new metadata and to make a system to support it which may include metadata handling and ID resolving functions. This data format GIST will be a foundation for the approaches.

## References


[1]**Burnard,L. and S.Bauman eds.** (2007) *The TEI Guidelines P5*, TEI

[2]**Ide, N and K.Sufdermanp** (2006) GrAF: A GrAF-based Format for Linguistic Annotations, Proc. of the Linguistic Annotation Workshop

[3]**ISO** (2004) *Date and time format*, ISO

[4]**ISO** (2006) *ISO/DIS 24610-1 Language Resource Management – Feature Structures – Part1: Feature Structure Representation*, ISO

[5]**ISO** (2011) *Language resource management – Feature structures – Part 2: Feature system declaration*, ISO

[6]**ISO** (2012) *Language resource management – Linguistic annotation framework (LAF)*, ISO

[7]**Ohya,K.** (2009) "Data Structure for Minority Language Corpora" (in Japanese), *IPSJ Symposium Series* Vol.2009, No.16, IPSJ

[8]**Ohya,K.** (2011) "Missing Services in Language Documentation in terms of Information Processing – The Report of LingDy Project –" (in Japanese), *IPSJ Symposium Series*, Vol.2011, No.8, IPSJ

[9]**Ohya,K.** (2012) "Corpus Sharing Strategy for Descriptive Linguistics" JADH2012

[10]LingDy; Linguistic Dynamics Science Project, `http://lingdy.aacore.jp/en/`